View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional de la Universidad de Alicante

Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 131-134 recibido 30-01-12 revisado 06-03-12 aceptado 18-03-12

EmotiBlog: A Model to Learn Subjective Information Detection in the New Textual Genres of the Web 2.0 -a Multilingual and Multi-Genre Approach-

EmotiBlog: un modelo para aprender a detectar la información subjetiva en los nuevos géneros textuales de la Web 2.0-una aproximación multilingue y multi género-

Ester Boldrini y Patricio Martínez –Barco
Universidad de Alicante- GPLSI
Carretera San Vicente del Raspeig s/n - 03690 San Vicente del Raspeig –Alicante
{eboldrini;patricio}@dlsi.ua.es

Resumen: Tesis doctoral con mención europea en procesamiento del lenguaje natural realizada en la Universidad de Alicante por Ester Boldrini bajo la dirección del Dr. Patricio Martínez-Barco. El acto de defensa de la tesis tuvo lugar en la Universidad de Alicante el 23 de enero de 2012 ante el tribunal formado por los doctores Manuel Palomar (Universidad de Alicante), Dr. Paloma Moreda (UA), Dr. Mariona Taulé (Universidad de Barcelona), Dr. Horacio Saggion (Universitat Pompeu Fabra) y Dr. Mike Thelwall (University of Wolverhampton). Califcación: Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Análisis de sentimentos, minería de opinions, nuevos géneros textuales, blogs, Web 2.0

Abstract: European Ph.D Thesis in Computational Linguistics, written at the Univeridad de Alicante by Ester Boldrini under the supervision of Dr. Patricio Martínez-Barco. The author was examined on January, 23th 2012 at the Universidad de Alicante by a commission composed by Dr. Manuel Palomar (Universidad de Alicante), Dr. Paloma Moreda (UA), Dr. Mariona Taulé (Universidad de Barcelona), Dr. Horacio Saggion (Universitat Pompeu Fabra) and Dr. Mike Thelwall (University of Wolverhampton). The mark obtained was *Sobresaliente Cum Laude.*
Keywords: Sentiment Analysis, Opinion Mining, New Textual genres, Blogs, Web 2.0

1 Introduction and motivation

This research has been focused on the EmotiBlog-Annotation-Model and the EmotiBlog-Corpus-Annotated, the multilingual and multi domain resource we created to detect subjectivity in the new textual genres with the intention of contributing to the improvement of the Sentiment Analysis task. Our main motivation was the huge amount of subjective data available on the Internet due to the wide employment of the new textual genres born with the Web 2.0. As it has been demonstrated, this data has an undeniable influence on people's behaviour and it can be exploited for many real-life applications. This is the reason why we concentrated on the Sentiment Analysis task, in charge of treating the opinionated data as first step to achieve robust applications, which are able to exploit this data. Our focus was on Sentiment Analysis on a multilingual level (English, Spanish and Italian) and we concentrated mainly in blogs, due to their demonstrated relevance in our society.

2 Contributions

The contributions we bring with the present work are numerous, but they can be summarised below.

First of all, we presented and analysed the relevance and usefulness of the new textual genres born with the Web 2.0 and above all blogs. After that, since blogs are producing a huge amount of data and they are considered as a consolidated textual genre, we decided to focus our research on this textual genre, describing in detail blogs main features,

ISSN 1135-5948
© 2012 Sociedad Española para el Procesamiento de Lenguaje Natural

peculiarities and the challenges they present regarding the language employed. Furthermore, we carried out an in-depth analysis of the state of the art in Sentiment Analysis in order to detect the aspects that still need improvement. Basing on the conclusions drawn from the previous research our conclusions are that: a) there is an evident scarcity of corpora in languages other than English, b) there is the need to collect corpora composed by blog posts, c) there is no fine-grained model for labelling the subjective data in the new textual genres and in a multilingual way. Starting from our state of the art conclusions we created a multilingual corpus of blog posts in English, Spanish and Italian about three topics: the Kyoto Protocol, the elections in Zimbabwe, the USA elections. After a deep empirical analysis of the corpus we collected, we built up the EmotiBlog-Annotation-Model, a fine-grained annotation scheme to detect subjectivity in these texts. In order to be the most adequate as possible to the needs of our corpus, we made an extensive analysis of the subjectivity definition and classification and choose the most suitable according to the needs of our dataset.The next step consisted in annotating part of our corpus and then perform a two-fold evaluation: intrinsic and extrinsic. The first one was focused on testing the model to check if it is easily employable by other annotators and if the annotation levels it contemplates are beneficial for the training of automatic systems to correctly classify the subjective data. The extrinsic evaluation was focused on checking if the information, annotated with the EmotiBlog could be exploited by systems to improve their performance of three Natural language Processing (NLP) tasks: Opinion Mining, Opinion Question Answering and Opinion Summarisation.

3 Intrinsic Evaluation

We carried out an intrinsic evaluation of the EmotiBlog-Corpus-Annotated and of the EmotiBlog-Annotation-Model. By calculating the inter-annotator agreement our aim was to check if the model was clear enough and if it allows for an easy and unambiguous annotation by different annotators. By means of performing the feature selection experiments we aimed at checking if the annotation allows a proper classification according to the EmotiBlog-Annotation-Model elements and attributes. We also measured the impact of the EmotiBlog-Annotation-Model in order to measure the effect of each element and refine the model.

4 Extrinsic Evaluation

The extrinsic evaluation was performed in order to check if the EmotiBlog-Corpus-Annotated was a useful and beneficial resource to improve the performance of the key Natural Language Processing tasks dealing with subjective data at the moment. At this stage, apart from using our resource, we also employed different other corpora and some of them of different textual genres to check also if the EmotiBlog-Annotation-Model would be suitable also for other textual genres. The NLP tasks in which we worked are listed below together with the EmotiBlog-Annotation-Model contribution:

4.1 Opinion Mining

We exploited the EmotiBlog-Corpus-Annotated in order to train the Machine Learning system to correctly classify the sentences of our documents depending on their subjectivity (subjective/objective) and intensity (low/medium/high) improving the baseline.

4.2 Opinion Question Answering

We employed our resource to help the system in learning how to classify the possible answers into positive and negative. The EmotiBlog-Annotation-Schema contemplates the necessary elements such as source, target, topic that are key aspects to improve the Opinionated Question Answering task, thus mixed with other resources can allow a considerable improvement of the baseline.

4.3 Opinion Summarisation

In this case the EmotiBlog-Annotation-Model was used to label our ad-hoc collection of blog posts about different topics and also to train the classification system to distinguish the sentence polarities, needed for the summary generation. The results obtained show that the discrimination has been done in a proper way and thus the approach is correct and the EmotiBlog-Annotation-Model together with the information it provides is a valid resource that allows the treatment and thus inclusion of the subjective information, thus providing a step

forward the state of the art, which has been treating objective data.

5 Most relevant results obtained

OPINION MINING IN SPANISH			
Classification using ten fold cross validation			
	Precision %	**Recall %**	**F1 %**
Neg.	89.2	96.9	92.9
Classification results using all n-grams and n-grams, n>2			
	Precision %	**Recall %**	**F1 %**
Negat.	92.3	96.2	94.2
Negat.	90.2	91	90.6
OPINION MINING IN ENGLISH			
Results for polarity and intensity classification using the models built from the EmotiBlog annotations			
Corpus	**Precision %**	**Recall %**	**F1 %**
JRC quotes II INTENSITY	36.4	51	42.9
JRC quotes II POLARITY	38.7	57.8	46.4
Results for emotion classification using the models built from the EmotiBlog annotations			
Corpus	**Precision %**	**Recall %**	**F1 %**
SemEval EmotiBlog Model I	29	18.9	22.9
QUESTION ANSWERING IN ENGLISH			
System		**F1 %**	
JIRS + SA+ET+LSA+SR (3 phrases)		0.6	
SUMMARISATION 10% COMPRESSION RATIO			
	Non Accept.%	**Understand %**	**Accept %**
Redun.	26	45	29
Gramm.	4	22	74
Focus	33	43	24

Tabla 1: Better results obtained

6 Thesis Overview

In Chapter 1 we made a deep study of our research framework. We presented the context of our research with the development of the Web 2.0 and the consequential growth of the new textual genres with all the implications they bring in our society. This represents a massive phenomenon and thus, due to the huge amount of subjective data available on the Web and the possible real time applications we can build up exploiting it, many are the disciplines that study how subjectivity is expressed. We described the main ideas of Neuroscience, Cognitive, Science, Psychology, but also Natural Language Processing, that is our research perspective. After that, we stressed on the importance of working with the new textual genres, we analysed the main ones and we justified why we decided to carry out our research mainly in blogs. Another consequence of the huge explosion of this new research area, many are the terms used interchangeably and thus in order to clarify our use of such terminology we defined the difference between Sentiment Analysis and Opinion Mining. According to our point of view the first one is the step that comes before and that allows a high performance Opinion Mining process. In fact, with Sentiment Analysis the language is properly analysed and treated before being exploited for concrete purposes. Apart from that, we also clarified what we mean when we employ the term subjectivity since this concept is strictly related to the EmotiBlog-Annotation-Model. In **Chapter 2** we carried out we analysed in detail the State of the Art with a special focus on the creation of linguistic resources for Sentiment Analysis. We classified them according to different criteria: size, language, textual genre and domain creating comparative tables in order to show the aspects we would like to improve with our work. The following step consisted of briefly presenting the main research carried out in the framework of Opinion Mining, Question Answering and Opinion Summarisation. **Chapter 3** represents the main nucleon of this work and our most important contribution to the improvement of the State of the Art presented previously. Here we described how we build up the EmotiBlog-Annotation-Model. We explained in detail the annotation levels EmotiBlog-Annotation-Model allows and we provided an in-depth description of the entire collection of the elements with their corresponding attributes providing examples in each case, thus entering in detail in the annotation process. Special emphasis was put on explaining why we chose such elements with their attributes and why we chose the values they have. **In Chapter 4** we carried out part of the intrinsic evaluation on the EmotiBlog-Annotation-Model by calculating the inter annotator agreement with the objective of checking if the annotation it allows was clear, unambiguous ad easy to perform. The test has been carried out with two experienced annotators and on the EmotiBlog-Kyoto

labelled in Spanish and we obtained positive results.The second part of the intrinsic evaluation of the EmotiBlog-Annotation-Model was carried out in **Chapter 5** in which we describe the results obtained in the feature classification experiments where we also we measure the impact of each element of the annotation model in order to understand which of them were beneficial for the classification purposes. After having checked the results, we refined the model producing its final version and the results obtained proved that our granularity approach is correct. Before performing the extrinsic evaluation of the EmotiBlog-Annotation-Model and EmotiBlog-Corpus-Annotated, in **Chapter 6** we described all the resources, tools and procedures that we will employ in the following chapters using the EmotiBlog–Corpus-Annotated for the improvement of the three Natural Language Processing Tasks we selected. In **Chapter 7** we start the extrinsic evaluation of the EmotiBlog-Annotation-Model and the EmotiBlog-Corpus-Annotated with the Opinion Mining Task. We exploited the EmotiBlog-Annotation-Model in order to train the Machine Learning system to train our system to correctly classify the sentences of our document depending on their subjectivity (subjective/objective), polarity (positive/negative) and intensity (low/medium/high) improving the baseline demonstrating that our resource is valuable for the classification process. Moreover, we entered in the task of Opinionated Question Answering using our resource to help the system to learn how to classify the possible answers into positive and negative. The EmotiBlog-Annotation-Schema has the necessary elements such as source, target, topic that are useful to improve the Opinionated Question Answering task, thus mixed with other resources can allow a considerable improvement of the baseline. Finally, we carried out the extrinsic evaluation of our resource in the framework of the Opinion Summarisation task. In this case EmotiBlog-Annotation-Model was employed to label our ad-hoc collection of blog posts about different topics and also to train the classification system to distinguish the sentence polarities, needed for the summary generation. The results obtained show that the discrimination has been done in a proper way and thus the approach is correct and EmotiBlog-Annotation-Model and the information it provides is a valid resource that allows high-level results.

7 Conclusions

As general conclusion, we can say that there is no doubt about the fact that Sentiment Analysis is an extremely challenging task but at the same moment it is a fascinating area of research. Even if much work has been done, there is still big room for improvement.

The intrinsic evaluation showed that the fine-granularity we chose is appropriate and the annotation seems to be feasible without any significant problem. Moreover, the EmotiBlog-Annotation-Model and Corpus-Annotated have also demonstrated to be effective for English, Spanish and Italian, but we also expect that it could work with high results in other languages that share a similar syntactic structure.

Concerning the textual genre, even if our collection is taken from blog posts, we also had the possibility during our experiments to work with other textual genres as newspaper articles and we proved that the EmotiBlog-Annotation-Model is compatible with other textual genres.

From the extrinsic evaluation we performed, we can assume that the EmotiBlog-Annotation-Model is beneficial in the three Natural Language tasks we selected. It is a key factor for the Opinion Mining task since it allows a better and more precise classification of sentence, element polarity and intensity.

The EmotiBlog-Annotation-Model employment is also strategic in the Question Answering task because it brings improvement of the baseline since it helps to correctly classify the answers and the possible snippets to retrieve the answers, but it also contemplates the key elements to improve the task (target, source, etc). We also concluded that our resource has a positive impact on the Opinionated Automatic Summarisation since it provides a module for analysing the subjectivity in sentences making the normal opinion summarisation task adapt for subjective content and thus, it means that EmotiBlog is valid and a key contribution also for this task.

References

Boldrini, E., Balahur, A., Martínez-Barco, P. Y Montoyo, A. (2012). Using EmotiBlog to Annotate and Analyse Subjectivity in New Textual Genres. Data Mining and Knowledge Discovery.DOI 10.1007/s10618-012-0259-9